Generation Adequacy: Who Decides?

While market-based methods for generation expansion are the preferred long-term approach, there is currently too much uncertainty regarding how consumers and suppliers will respond to spot price signals. Therefore, policymakers should encourage demand-side experiments and investments to ensure that, when prices rise, customers will be able to respond.

Eric Hirst and Stan Hadley

Eric Hirst *is a consultant in electric-industry restructuring in Oak Ridge, TN. His work focuses on bulk-power issues related to reliability, ancillary services, generation and transmission adequacy, and the design of competitive regional transmission organization (RTO) markets, as well as stranded costs.*
Stan Hadley *is a research engineer at the Oak Ridge National Laboratory, where he developed the ORCED dispatch model. Using it, he has studied policy options for greenhouse gas reductions, regional power prices under restructuring, stranded cost recovery, and energy efficiency impacts on air emissions.*
The work reported here was sponsored by the Edison Electric Institute.

I. Background

Historically, decisions on the amounts, locations, types, and timing of investments in new generation have been made by vertically integrated utilities with approval from state public utility commissions. As the U.S. electricity industry is restructured, these decisions are being fragmented and dispersed among a variety of entities. As generation is deregulated and becomes increasingly competitive, decisions on whether to build new generators and to retire, maintain, or repower existing units will increasingly be made by unregulated for-profit entities. These decisions will be based largely on investor assessments of future profitability and only secondarily on regional reliability requirements. In addition, some customers will choose to face real-time (spot) prices and will respond to the occasionally very high prices by reducing electricity use at those times. During the transition from a vertically integrated, regulated industry to a deintegrated, competitive industry, government regulators and system operators may continue to impose minimum-installed-capacity requirements on load-serving entities. As the industry gains experience with customer responses to real-time pricing, these requirements will likely disappear.

As part of a project for the Edison Electric Institute, we examined the commercial and reliability aspects of investments in new generation.[1] This article reviews historical data and projections on new generating capacity, discusses the pros and cons of alternative ways to maintain adequacy, and quantifies the effects of mandating minimum planning-reserve margins versus reliance on market prices to stimulate investments in new generation.

Figure 1 shows utility forecasts of generation-capacity margins from 1990 through 1998 and projections through 2007.[2] Nationwide, reserve margins declined from 22 percent in 1990 to 16 percent in 1997, and are expected to decline further to 10 percent in 2007.

For at least the past several years, generation adequacy has declined. Utility reports to the North American Electric Reliability Council (NERC) suggest that these trends will continue. Indeed, the latest NERC reliability assessment is more pessimistic than earlier ones, primarily because of utility reluctance to build new generation because of uncertainties about cost recovery for such investments, loss of integration between generation and transmission planning, reluctance of independent power producers (IPPs) to reveal their generation plans much in advance of actual construction, possible double-counting of some generating capacity as more suppliers rely on purchases from other entities, and uncertainty over the extent to which demand-side responses will reduce the need for new generation. On the other hand, IPPs plan to build 69,000 MW of new capacity.

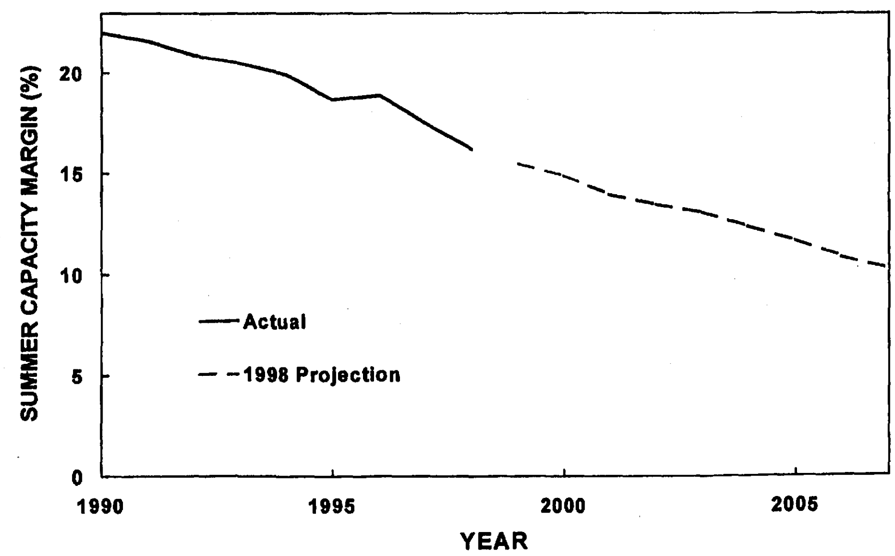

Figure 1: U.S. Summer Generation-Capacity Margins from 1990 through 2007

II. Concepts

Historically, utilities maintained "extra" generating resources for short- and long-term purposes; this article focuses on long-term reserves, often called planning reserves, and does not deal with operating reserves. At least two mechanisms can be used to maintain generation adequacy:

- Reliance on markets—the interactions of consumers and suppliers acting through the mechanism of volatile spot prices—to decide what types of generation to build and when and how much electricity to consume. California adopted this approach.
- Reliance on the traditional system of having a central agency (e.g., the independent system operator or state regulator) specify an appropriate minimum reserve margin based on estimates of the value of lost load (VOLL) and other factors (e.g., forced and planned outage rates for different types of generating units). This reserve margin is then imposed on all load-serving entities (LSEs). The three Northeastern ISOs (Pennsylvania-New Jersey-Maryland, New York, and New England), all of which developed from traditional tight power pools, use this approach.

The United Kingdom uses a third system. There, the National Grid Company calculates, on a day-ahead basis, the expected loss-of-load probability (LOLP) for each 30-minute period. This LOLP is then multiplied by the assumed VOLL of about $4 per kWh to develop a capacity charge, which is added to the system marginal price. This approach has received little attention in the United States, perhaps because the capacity charge is too easy to manipulate for companies that own large

1040-6190/99/$–see front matter PII S1040-6190(99)00073-1

amounts of generation. They can declare units unavailable in the day-ahead market and then make them available in real time to collect the high-capacity charge caused by the unavailability declarations.

Thus, the key issue on generation adequacy is whether (1) competitive generation markets for capacity and energy will be sufficient to maintain societally desirable levels of reliability, or (2) government regulators and system operators will need to impose mandatory minimum-reserve obligations on LSEs to ensure that customers are not *in*voluntarily interrupted from their electricity supplies.

These two options should produce different outcomes in

- Hourly energy prices, with reliance on real-time markets likely to yield lower average prices and costs but greater price volatility;
- Customer load shapes, with reliance on real-time markets likely to yield higher load factors; and
- Generation portfolios, with reliance on real-time markets likely to yield relatively more baseload capacity.

III. Discussion

Our review of the literature as well as discussions with several market participants yielded surprising agreement. Almost everything we read and everyone we spoke with believes that—in the long run—generation adequacy will be left to markets, with little involvement by government regulators. To do otherwise, most people recognize, would interfere with the workings of competitive energy markets. That is, the energy and capacity markets are closely coupled.

On the other hand, most people agreed that we may need a multiyear transition period while suppliers and, especially, retail customers learn how to respond appropriately to rapidly changing (e.g., hourly) electricity prices. (We first have to permit retail customers to face these time-varying prices; in most parts of the country, customers still face prices based on embedded costs that are largely time-invariant.) During this transition period, prudence may require mandated planning-reserve margins.

Almost everyone believes that generation adequacy will be left to markets, with little involvement by regulators.

Proponents of market-based decisions on generation retirements and expansions worry, however, that electricity price spikes, such as the one that occurred in the Midwest in June 1998, will bring forth inappropriate government price controls. (Prices spiked, in part, because almost all retail customers paid only traditional, embedded-cost rates and did not face these very high wholesale prices.) According to one analysis of the repercussions of the spike,

> Market fluctuations heighten regulatory risk. The jury is still out on whether policy markets (legislators and regulators—elected and appointed officials) and the public can tolerate price fluctuations in the energy market. After the [June 1998 Midwest] price spike, industrial consumers, utilities, legislators, and others called for price caps or price regulation to limit prices on the upside. (No consumers or legislators have clamored for price floors to limit producers' losses during shoulder seasons when prices are microscopic.) So far, FERC and the Congress have resisted the call for price caps. However, in the future, additional price anomalies, even for brief periods, will reduce regulators' and politicians' enthusiasm for a competitive electricity commodity market.[3]

In support of the market option, two students of the phenomenon noted that,

> Price spikes . . . provide market participants with important information needed for trading and capacity investment decisions. Price increases signal price-sensitive customers that it is time to conserve, and they tell producers that it may be time to expand capacity. Price increases also give producers and consumers incentives to change their behavior in ways that mitigate severe spikes; producers can profit by investing in new capacity, and consumers can make themselves better off by reducing peak period demand. . . . It is true that only some customers need to moderate their usage to reduce peak prices for everyone. But in the absence of a competitive market, we have no way of

PII S1040-6190(99)00073-1

knowing which customers are most willing to do this.[4]

Figure 2 schematically illustrates the supply/demand balances with and without an explicit installed-capacity requirement. The dashed line that slopes up to the left represents consumer demand, and the stairstep line that slopes up to the right represents generating capacity. With a reserve-margin requirement of 11,500 MW, supply and demand equilibrate at a price equal to the variable cost (fuel plus variable operations and maintenance, or O&M) of the last (marginal) unit online at that time. If, however, there is no required reserve margin and market forces yield only 10,500 MW of available capacity, the price of electricity will rise above the variable cost of the last unit online when unconstrained demand exceeds 10,500 MW. The amount of price increase (the pure capacity price in Figure 2) is a function of the demand elasticity for electricity. The more responsive customer demand is to changing electricity prices, the smaller this capacity price will be.

This example makes two points. First, even if there is "insufficient" capacity from an engineering perspective, price-responsive demand and supply will equilibrate, and the bulk-power system will not crash. This equilibrium occurs because some customers would rather forego some consumption than pay the high price associated with this situation. Second, at times of high demand, spot prices will be higher if there is no required reserve margin. In other words, specifying a minimum amount of installed generating capacity will suppress spot prices at certain times. Economists argue that this suppression of a valuable price signal will undercut energy and capacity markets.

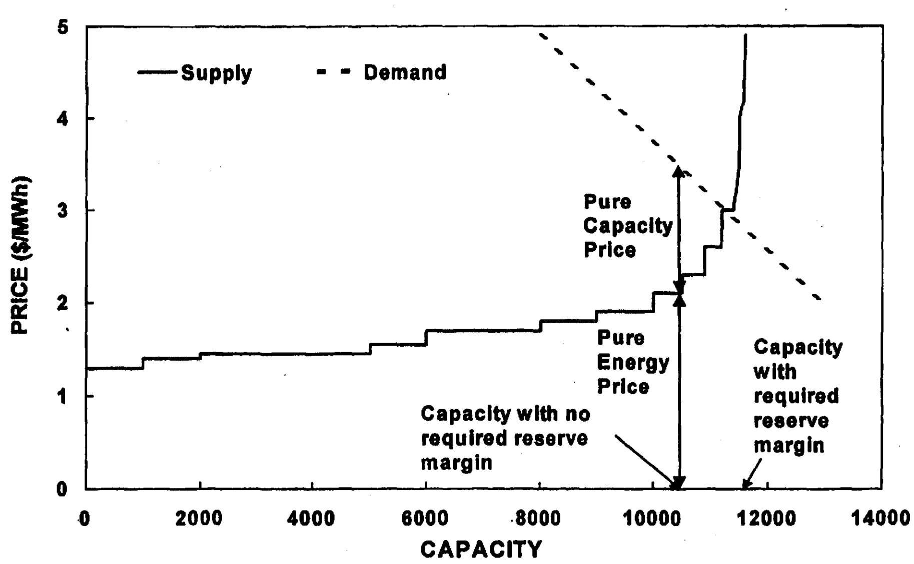

Figure 2: With an Installed-Capacity Requirement of 11,500 MW, Supply and Demand Balance at a price of 3.0 cents/kWh. With No Capacity Requirement and Only 10,500 MW Online, Unconstrained Demand Would Exceed Supply, and Prices Would Rise to 3.4 cents/kWh.

Requiring a minimum reserve margin creates two markets (installed capacity and energy) with no assurance that they will be in equilibrium, according to F.C. Graves and others who have studied the matter.[5] This requirement will suppress energy prices and undercut demand-side participation in reliability. Requiring minimum reserve margins and the associated two-part pricing for capacity and energy "will undermine the benefits of power industry restructuring." On the other hand, energy-only markets "will induce efficient capacity planning—which has been the real problem in the past (not inefficient dispatch) and which is where the real opportunities for future efficiency gains lie. It will also encourage demand-side participation in peaking reserves, and forward contracting for risk protection . . ."

Graves *et al.* note several problems with the traditional engineering approach to maintaining generation adequacy:

- Setting fixed capacity requirements to deal with what is inherently a very uncertain situation. The uncertainties deal with the timing, extent, and duration of forced outages and with the tremendous variation among customers in their value of lost load.
- Static demand curves with zero price elasticity in spite of the evidence from real-time pricing programs that customers differ substantially in their willingness and ability to respond to changing electricity prices.

• The assumption that forced outages occur in a completely random fashion. In competitive markets, generation owners will work hard to assure that their plants are available during high-price periods.

• With customer choice, customers may want to choose their level of reliability, not have it specified for them by a central authority.

L.E. Ruff, examining the matter at National Economic Research Associates, cautions that reliance on market prices can work well only when spot prices accurately reflect costs: "In practice, hourly energy prices tend to be below the average of the 'correct' instantaneous prices over the hour, particularly during critical hours when peaking capacity is needed. This is because the ISO must use a relatively simple, mechanical rule to determine a single energy price for the hour . . . and most simple pricing rules miss within-hour effects."[6]

Loehr (1998) offers an opposing, cautionary view of generation adequacy:

> The argument in favor of deregulation is that investors will build new units in response to the demands of the market. Perhaps so. But there are some major concerns. For most electric power systems in the United States, the actual load exceeds 90 percent of the peak load only 1 to 2 percent of the time. In the past, utilities had an "obligation to serve" all of the load all of the time; even the last 10 percent. This was part of the regulatory compact. Thus they planned, built and operated as much generation as was required by the peak load. But today there is a real question as to whether, in an industry driven by competition and the marketplace, investors will be willing to commit financial resources to supply customer load which will be realized only a few hours a year. As far as actual or potential generation owners are concerned, this is a basic question of price and price signal.[7]

Loehr appears to suggest that, in a competitive electricity industry, generation owners cannot earn a profit building plants that operate only 1 to 2 percent of the time; given a choice, they would not build such plants. Therefore, society needs to make them do so through minimum-reserve requirements. Society then requires all electricity consumers, regardless of how highly they value electricity consumption at times of tight supplies, to pay for this capacity. Loehr appears not to consider the possibility that these plants will be unprofitable because consumers would rather reduce consumption than pay such high prices. The economists argue that this "extra" generating capacity should be built only if customers are willing to pay the very high spot prices associated with the very infrequent use of these units; otherwise, enough customers will reduce their demand sufficiently to yield a supply/demand balance at a lower level of generating resources and lower peak-period prices.

Requiring a minimum reserve margin creates two markets, with no assurance that they will be in equilibrium.

Jaffe and Felder believe that mandated capacity requirements are needed because such capacity benefits society at large, not just the owners of such capacity.[8] Such societal benefits are especially large for electricity because of its pivotal role in modern society, the real-time nature of electricity production and consumption (which occur within milliseconds of each other), and the difficulty of storing electricity.[9] They note that policymakers can either set minimum-reserve margins or subsidize capacity with an up-front dollars-per-kW-year payment for capacity. In principle, the two approaches should yield the same outcome.

NERC raises concerns that "few, if any, customers understand the implications of contracting for other than firm power supplies and firm transmission services."[10] Because of the long tradition of ample supplies and the use of interruptible rates to offer implicit discounts to large industrial customers, these customers are used to very few interruptions in service. Indeed, industrial customers, when interrupted, often are angry. Thus, it is an open question how customers will respond to real-time pricing. In addition, only a few electric utilities (e.g., Georgia Power) have much experience and a clear understanding of whether and how customers might respond to real-time pricing.

IV. ISO Approaches

Bulk-power operations in California are split between the Power Exchange (PX) and the independent system operator (ISO). The PX runs day-ahead and day-of energy markets. In addition, the ISO operates a real-time energy market to balance generation and load during each hour. Neither the PX nor the ISO specifies installed-capacity requirements for market participants, and neither entity operates an installed-capacity market.

Between April 1, 1998, and March 31, 1999, the weighted average price of electricity in the PX day-ahead market was $26.6 per MWh. For 12 percent of the hours, prices were at or below $10 per MWh. At the other extreme, prices were at or above $100 per MWh for 1.1 percent of the hours, with prices ranging as high as $200 per MWh; these prices are well above the marginal costs of the most expensive units in California and reflect the pure capacity price shown in Figure 2.

The California ISO points to the number and size of the proposed power plants in California (16 projects with a total capacity of more than 10,000 MW as of spring 1999) as evidence that competitive markets for capacity can work.[11]

Nationwide, as of October 1998, developers had announced plans to build 109 merchant plants with a total generating capacity of over 56,000 MW.[12] (According to the Electric Power Supply Association, the total capacity of announced merchant plants had increased to almost 69,000 MW by February 1999.) Although many of these plants will likely not be built because of problems with siting, state approval, financing, or transmission access, these plans suggest that regulatory mandates are not needed to bring forth new generating capacity.

New England's two capacity markets in addition to the energy and ancillary services markets may be historical artifacts.

The PJM *Reliability Assurance Agreement* (RAA) establishes the obligations of all LSEs within the PJM control area to provide the amount of installed generating capacity that PJM determines is needed to maintain reliability.[13] The PJM Reliability Committee determines the forecast pool requirement, the reserve margin for the PJM Control Area required as part of this agreement. The RAA "is intended to ensure that adequate Capacity Resources will be planned and made available to provide reliable service to loads within the PJM Control Area, to assist other Parties during Emergencies, and to coordinate planning of Capacity Resources consistent with the Reliability Principles and Standards."[14]

The PJM Reliability Committee determines the forecast pool requirement for capacity resources using "probability methods" and establishes criteria for use of capacity resources during emergencies. The forecast pool requirement is intended to "ensure a sufficient amount of capacity to meet the forecast load plus reserves adequate to provide for the unavailability of Capacity Resources, load forecasting uncertainty, and planned and maintenance outages."[15] The focus is on the peak season, which for PJM overall is the summer.

In October 1998, PJM established the monthly PJM Capacity Credit Markets to allow PJM market participants to buy and sell capacity credits to meet their obligations under the RAA.[16] The markets for January through May 1999 cleared 5,000 MW at prices that ranged from $50 to $80 per MW-day. The average price during this five-month period was $65 per MW-day (equivalent to $24 per kW-year, assuming that capacity is valued equally for every day of the year). PJM began daily markets (conducted a day ahead) in installed capacity in January 1999. For the four months of January through April 1999, the daily price averaged less than $5 per MW-day, far below the prices in the monthly markets. (It is not known whether these price variations reflect seasonal differences, differences between daily and monthly markets, or lack of familiarity with these new products.)

  PII S1040-6190(99)00073-1

The ISO New England (1997) approach differs from the PJM and New York approaches in that New England has two capacity components: monthly installed capability and hourly operable capability.[17] Operable capability refers to "any generating unit or units in any hour . . . which is operating or available to respond within an appropriate period to the System Operator's call to meet the Energy and/or Operating Reserve and/or AGC [automatic generation control] requirements of the NEPOOL [New England Power Pool] Control Area. . . ." New England market participants are *required* to bid all their operable capacity in excess of their obligations into the hourly operable capability market.

It is unclear why New England requires two capacity markets in addition to the energy and ancillary-services markets. Our discussions with several people in the region suggest that the two markets are historical artifacts and that, within a few years, one or both will be eliminated.

ISO New England noted that "NEPOOL has had an ICAP [installed-capability] requirement since its inception. This requirement has been important in maintaining reliability in New England for over 25 years."[18] These comments were in response to criticism from Cramton and Wilson, who reviewed New England's proposed market rules. Specifically, they wrote about the installed-capability requirement:

> This holdover from an era of regulation is unique in the electricity industry, which is the only one that does not expect suppliers to cover fixed costs, such as capital and maintenance, from the market price of its output. . . . The capacity markets are a holdover from the regulated setting, when capacity decisions were not made in response to price expectations. In the transition to a competitive market, the capacity markets may serve a useful role in coordinating investments in capacity. However, once competitive electricity markets are established in New England, it would be appropriate for the capacity markets to terminate.[19]

ORCED analyses suggest that centralized decisions concerning the amount of generating capacity to maintain may often be wrong.

New England may maintain both installed- and operable-capability requirements because the installed-capability requirements are largely independent of availability. The installed-capability requirements relate primarily to "iron in the ground" without regard to the ability of that unit to operate any time soon. For example, the three large Millstone nuclear units were out of service for 18 months or longer, during which time they continued to qualify as installed capability in New England.

This discussion of the need for installed- and operable-capability markets raises the difficulty in determining what to include as installed capacity. Over what time period should generating-unit availability be measured? Should availability be determined on a daily, monthly, seasonal, or annual basis? Because the need for capacity is generally greatest during winter and summer peak periods, it may be most important to measure availability during those time periods.

V. Quantitative Analysis of Generation Adequacy

We used the Oak Ridge Competitive Electricity Dispatch model (ORCED) to analyze and quantify these generation-adequacy issues.[20] ORCED is a simple strategic planning model that simulates the operations of, and resultant prices and producer profits from, competitive bulk-power systems.

The ORCED version used here consists of one region (within which transmission is unconstrained) with 50 generating units. Each generating unit is characterized by its heat rate, fuel type and cost, variable and fixed O&M costs, and capital costs. The model distinguishes between plants already online, for which capital costs are sunk, and plants that have not yet been built, for which capital costs are fully avoidable.

Consumer loads are represented by two load-duration curves for offpeak and onpeak seasons. Consumer responses to changes in electricity prices are represented by three input demand elasticities: (1) an unserved-energy elasticity that adjusts demand

PII S1040-6190(99)00073-1

down during those periods when unconstrained demand would otherwise exceed capacity, used to calculate the market price at which supply and demand equilibrate; (2) a time-of-use elasticity that changes the shape of the load-duration curve in response to changes in hourly spot prices; and (3) an overall elasticity that adjusts annual consumption up or down on the basis of decreases or increases in overall electricity price. This analysis uses only the first of these three elasticities.

ORCED is both an optimization model and a simulation model. We used ORCED to minimize the avoidable cost of electricity production for a given year subject to various constraints. The constraints limited the megawatt capacity of new generation that could be constructed and ensured that no generating units lost money (i.e., every unit earned revenues at least equal to the unit's avoidable costs). Given the set of generating units online, the simulation portion of the model dispatched these units in least-cost order for the analysis year.

We ran two sets of cases with ORCED. In the first set, we fixed the unserved energy elasticity at 0.05[21] and ran several cases with different values of reserve margin. Because ORCED deals with energy and not with ancillary services, these reserve margins should be increased by at least 5 percentage points to reflect the need for generating capacity for regulation, spinning reserve, and supplemental reserve.[22] We set up these model runs to minimize the avoided cost of electricity production, taking into account both the construction of new generators and the operation of the existing fleet. We then added an annual capacity payment (in dollars per kW-year) to ensure that the most unprofitable unit just broke even. This capacity payment was determined by dividing the monetary losses for each generator (for those generators that lost money) by the availability-adjusted capacity of each generator. Given the required amount of installed

capacity, the payment was then set equal to the highest dollar-per-kW loss to ensure that no generator lost money.[23] This capacity payment (which cannot exceed the carrying cost of a new combustion turbine, around $60 per kW-year) is then added to the price of electricity that consumers pay.

In the second set of cases, we varied the unserved-energy price elasticity and let the model determine the "optimal" reserve margin. We calculated an O&M adder (expressed in dollars per kW-year) for those units that operate less than 10 percent of the hours. We added this factor (which ranged between 0 and about $2 per kW-year) to ensure that plants operating for only a few hours a year would recover their avoidable fixed costs. (ORCED converts the adder to an energy-price premium paid to those units; the premium increases as capacity factor declines below 10 percent.) The rationale for including this O&M adder is the same as that used to justify inclusion of the capacity adder in the fixed-reserve margin cases—to guarantee that no generator loses money.

Our base case is an electric system with peak demand of 7,000 MW and annual energy consumption of 38,600 GWh, yielding a system load factor of 63 percent. With an unserved-energy elasticity of 0.05 and a reserve margin of 5 percent, the shares of generating capacity and energy are: coal (39 and 50 percent), gas (33 and 25 percent), nuclear (13 and 17 percent), hydro (11 and 8 percent), and oil (4 and 0 percent). All the new generating units (960 MW) are gas-fired combined-cycle units. The annual average price of electricity is 2.89 cents per kWh (ranging from 0.5 to 22.8 cents per kWh throughout the year).

We ran cases with reserve margins ranging from 1 to 15 percent (**Figure 3**). Model results show that the total cost of electricity production increases at very low and at very high levels of reserve margins. The cost curve has a broad minimum at around a 5 percent reserve margin. As the required reserve margin increases beyond the optimum value, the total price of electricity (i.e., the spot price of energy plus the capacity payment) increases. However, the spot price of energy declines and the capacity

  PII S1040-6190(99)00073-1

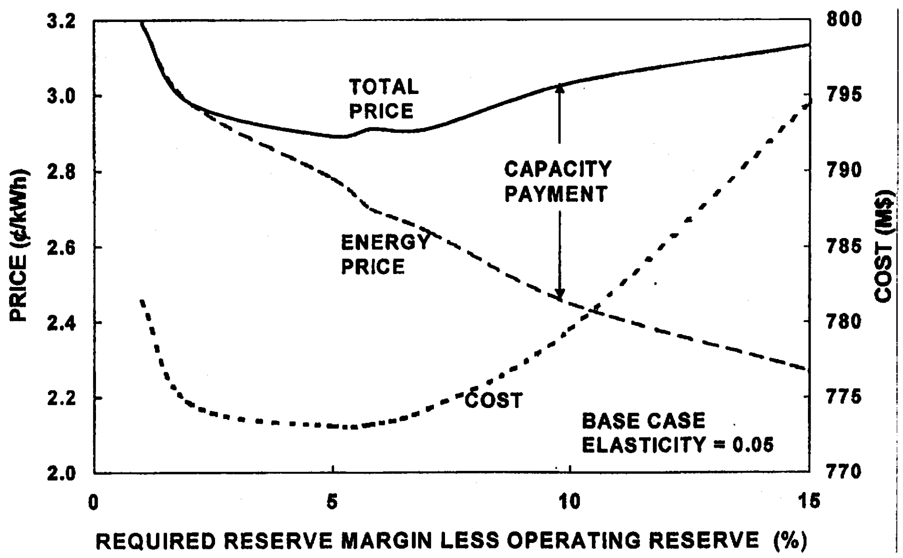

Figure 3: ORCED Results Showing Electricity Prices and Costs as Functions of Required Reserve Margins

charge increases as the required reserve margin increases. Thus, setting the reserve margin too high drives a substantial wedge between energy prices and the total price that consumers see, undercutting the operation of the energy market. In addition, higher specified reserve margins lower peak period prices and the volatility of prices during the year. As the required reserve margin is increased from 1 percent to 5 percent to 10 percent, the maximum spot price declines from 28 cents per kWh to 23 cents per kWh to 19 cents per kWh.

Electricity costs and prices are high at very low values of reserve margin because the amount and cost of unserved energy are high. At the other end of the spectrum, prices and costs are high because of all the "excess" capacity that must be supported through the capacity payment.

As the specified reserve margin is increased, the amount of new generating capacity brought online increases. For these cases, all the new capacity is gas-fired. Because this new capacity is very efficient, it accounts for larger shares of energy production than of generating capacity. For example, at the "optimal" 5 percent reserve margin, new gas-fired generation accounts for 14 percent of generating capacity and 19 percent of energy production. For reserve margins of 10 percent or below, all the new capacity is combined-cycle units; at higher values of reserve margins, some of the new capacity is combustion turbines. At high values, the capacity factor for generation declines. For low-capacity-factor operation, combustion turbines are a more economical choice than combined-cycle units because of their low capital costs (offset by their higher operating costs).

We then ran ORCED with no required reserve margin but with different values for the consumer response to price changes when unconstrained demand would otherwise exceed supply (the first elasticity factor discussed above). The results show substantial benefits from encouraging consumer response to high prices (**Figure 4**). An increase in elasticity from 0.01 to 0.04 cuts costs and prices by 7 percent and 15 percent, respectively. The results also show that, beyond an elasticity of about 0.04, there is little additional benefit to greater consumer response to price changes. At higher levels of elasticity, costs and prices decline very slowly. As demand elasticity increases, the maximum spot price and price volatility decrease. As the elasticity increases from 0.02 to 0.05 to 0.10, the maximum spot price declines from 339 cents per kWh to 38 cents per kWh to 13 cents per kWh.

As expected, the reserve margins chosen by ORCED as a function of demand elasticity decrease as elasticity increases. The reserve margin is about 5 percent at an elasticity of 0.05, consistent with the results shown in Figure 3. The price of unserved energy—the market price of electricity at times when supplies are not sufficient to meet *unconstrained* demand—declines steeply as elasticity increases. However, at very low values of elasticity, the price can be quite high; at an elasticity of 0.015, the price of electricity would have to be 629 cents per kWh to ensure that demand declines enough to match supply.

What can we conclude concerning the two primary approaches to ensuring that enough generating capacity is available so that customers will not be *in*voluntarily disconnected from the grid? Those most concerned about reliability note that requiring minimum planning reserves (1) ensures that "enough" generation will exist; (2) uses an approach that worked well in the past; (3) reduces the volatility of electricity prices; (4) protects customers, most of whom do not want to deal with the complexities of time-varying prices, from such volatility; and (5) ensures that the positive externalities associated with extra generating capacity are maintained. Those most concerned about economic efficiency and development of competitive markets for electricity counter that (1) the amount of generating capacity that is "enough" depends on customer response to prices, which varies across customers and customer classes; (2) because an approach worked well in a vertically integrated, monopoly-franchise industry is no assurance that it will work well in a deintegrated and competitive industry; (3) price volatility sends important economic signals to consumers and producers concerning when and how much electricity to consume and produce; (4) only a small fraction of loads needs to be price-sensitive to equilibrate demand and supply and to eliminate the need for mandated planning reserves; and (5) no public benefits are associated with generation adequacy beyond the private benefits.

In addition, the market proponents favor market decisions on generation investment because it (1) lets markets make both capital and operating decisions, which is why we are creating competitive electricity markets in the first place; (2) encourages customer participation in the provision of reliability services; and (3) will yield lower average electricity prices and costs to consumers. The reliability proponents respond that: (1) reliability is in part a public good that cannot be left entirely to the self-interests of market participants; (2) there is too much uncertainty about whether, when, and by how much customers will reduce demand in response to high spot prices to consider demand-side responses a resource for reliability; and (3) the costs of major outages are so high that they wipe out any savings associated with lower electricity prices.

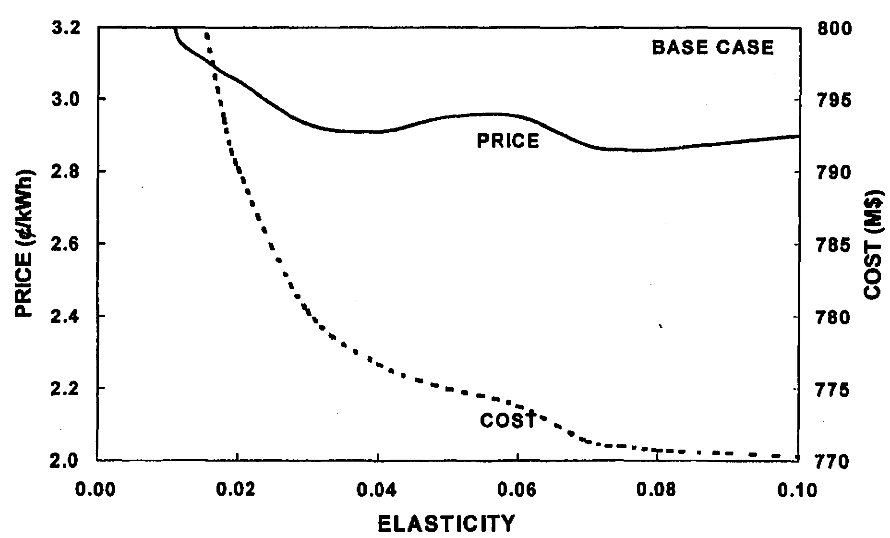

Figure 4: ORCED Results Showing Electricity Prices and Costs as Functions of the Unserved-Energy Price Elasticity

The ORCED analyses suggest that centralized decisions concerning the amount of generating capacity to maintain may often be wrong, yielding either too much capacity or not enough. So many factors affect the "optimal" amount of generating capacity (e.g., the prices of fossil fuels, the amount and types of generating capacity already online, and customer load shapes and price elasticities) that it is difficult for central planners to make the right choice. However, the ORCED results displayed in Figure 3 show a broad optimum, ranging in this case over several percentage points around the true optimum.

On the other hand, relying on the actions of consumers and suppliers in response to time-varying spot prices works well only if consumers can and do respond to high prices. Therefore, electricity policymakers should encourage demand-side experiments and investments to ensure that, when prices rise, customers will be able to respond.

VI. Summary and Conclusions

The long, awkward, and difficult transition from a highly regulated, retail-monopoly-franchise structure to a competitive structure creates generation-adequacy problems. Perhaps the key problem is the absence of a demand-side response to real-time pricing. Economic theory suggests that con-

  PII S1040-6190(99)00073-1

sumers and suppliers, in response to real-time prices, will take appropriate steps to ensure generation adequacy. But, if most retail consumers continue to face traditional tariff prices that have little or no temporal variation, this approach will be short-circuited. Until real-time pricing is available to at least some retail customers, traditional approaches to maintaining generation adequacy may be needed.

Our key findings and conclusions are as follows:

- Generation-capacity margins have been declining for at least a decade and are expected to continue to decline.
- Whether the decline in reserve margins reflects increased productivity or shortfalls in reliability is not clear. It is clear, however, that the transitional state of the U.S. electricity industry (half competitive and half regulated) leads to tremendous uncertainty, which may limit investments in long-lived assets, such as generating units.
- Generation adequacy could be maintained in competitive electricity markets through: (1) sole reliance on markets, acting through time-varying spot prices, or (2) continuation of the historical practice of setting minimum requirements on installed capacity that must be met by all load-serving entities.
- Market-based methods for generation expansion seem, both to us and to most of the people we talked with, the preferred long-term approach. ■

Endnotes:

1. E. Hirst, B. Kirby, and S. Hadley, *Generation and Transmission Adequacy in a Restructuring U.S. Electricity Industry* (Edison Electric Institute, Washington, DC), June 1999.

2. North American Electric Reliability Council (NERC), *Reliability Assessment: 1998–2007* (Princeton, NJ), Sept. 1998.

3. E. Lapson, R. Hornick, S.M. Fetter, and R. Pellecchia, *Electricity Price Spike: Lessons Learned* (Fitch IBCA, New York), Oct. 29, 1998.

4. R.J. Michaels and J. Ellig, *Electricity Passes the Market Test: Price Spikes in the Summer of 1998* (Mercatus Center, George Mason University, Fairfax, VA), Oct. 1998.

5. F.C. Graves, E.G. Read, P.Q. Hanser, and R.L. Earle, *One-Part Markets for Electric Power: Ensuring the Benefits of Competition* (The Brattle Group, Cambridge, MA), Oct. 1998.

6. L.E. Ruff, *Competitive Electricity Markets: Why They are Working and How to Improve Them* (National Economic Research Associates, Los Angeles), May 12, 1999.

7. G. Loehr, *Transmission Reliability: 'Brave New World' or 'Brave New World?'* EUCI Conference on Reliability and Competition, Denver, CO, Nov. 3, 1998.

8. A.B. Jaffe and F.A. Felder, *Should Electricity Markets Have a Capacity Requirement: If So, How Should It Be Priced?* PROCEEDINGS: 1996 EPRI CONFERENCE ON INNOVATIVE APPROACHES TO ELECTRICITY PRICING (EPRI TR-106232, Electric Power Research Institute, Palo Alto, CA), March 1996.

9. These societal benefits might include avoidance of the looting and violence that can erupt during a major blackout and the maintenance of electrical service to vital societal functions, such as hospitals, police and fire stations, traffic lights, and airport traffic-control systems.

10. NERC, *supra* note 2.

11. California Independent System Operator, *1998 Transmission Reliability Report* (Folsom, CA), July 1998.

12. C.W. Thurston, *Merchant Power: Promise or Reality?* PUB. UTIL. FORTNIGHTLY, Jan. 1, 1999, at 15–19.

13. PJM Interconnection, *Reliability Assurance Agreement among Load Serving Entities in the PJM Control Area* (Norristown, PA), Sept. 15, 1998.

14. *Id.*

15. *Id.*

16. PJM Interconnection, *Schedule 11: PJM Capacity Credit Markets* (Norristown, PA), Oct. 14, 1998.

17. ISO New England, *ISO New England Assessment of the Competitiveness and Effectiveness of the NEPOOL Markets, Exhibit A* (Holyoke, MA), Sept. 1998. The difference between installed and operable capability appears to be inoperable capability. Because no one should want to purchase inoperable capability, these two markets may be redundant. In addition, operable capacity seems to duplicate the real-power ancillary services.

18. *Id.*

19. P. Cramton and R. Wilson, *A Review of ISO New England's Proposed Market Rules* (ISO New England, Holyoke, MA), Sept. 9, 1998.

20. S. Hadley and E. Hirst, *ORCED: A Model to Simulate the Operations and Costs of Bulk-Power Markets* (ORNL/CON-464, Oak Ridge National Laboratory, Oak Ridge, TN), June 1998.

21. An unserved-energy elasticity of 0.05 (−0.05, to be precise) means that a 1 percent increase in the price of electricity cuts demand by 0.05 percent. This elasticity determines how high the spot price of electricity must rise to reduce demand so that demand equals online supply.

22. E. Hirst and B. Kirby, *Unbundling Generation and Transmission Services for Competitive Electricity Markets: Examining Ancillary Services* (NRRI 98-05, The National Regulatory Research Institute, Columbus, OH), Jan. 1998.

23. This no-loss constraint is essential in competitive electricity markets. Were a unit to lose money continuously, it would go out of business, which would reduce the amount of installed capacity below the minimum specified. This problem is not solved by a change of ownership because it is a function only of operating, not capital, costs.